A Visual Tracking Algorithm in Large-Scale Video with Convolutional Neural Networks

Ying Wang¹,²*, Xiaoju Ning¹

ABSTRACT
Convolutional Neural Networks (CNNs) had become a powerful model for solving many problems. In this paper, a novel visual tracking algorithm in large scale video based a trained CNN is proposed. The algorithm can track the trajectory of a moving object in a video with complex background quite precisely. Different from the most existing algorithms, we offline pre-trained a CNN through massive images data to obtain generic image features, which can be used the online tracking process. The trained CNN consists of shared layers and multiple branches of domain-specific layers, each branch is used for classification to identify target in each domain. When tracking a moving object in a new video sequence, a new network by combining the shared layers in the pre-trained CNN with a new classification layer is constructed. Experiment results show that performance of the proposed algorithm is excellent for some representative tracking benchmarks.

Key Words: Convolutional Neural Networks, Visual Tracking, Classification Layer, Offline Training

DOI Number: 10.14704/nq.2018.16.6.1664

Introduction
Visual tracking is a widely used basic computer vision task (Avidan, 2007). Although there had been lots of improvements in the recent year (Cai et al., 2017; Wang et al., 2016; Kumar et al., 2015), there exist enormous challenges in designing an excellent tracker which can handle significant appearance modeling and background separating very well.

In many problems of visual tracking, object appearance modeling is one of the most important problems. Many methods have been used to maintain the adaptability and robustness of the model (Xie et al., 2012; Kalal et al., 2012; Shi et al., 2014). Most existing tracking algorithms assume that the appearance of the target object changes smoothly with time. However, this strategy may not be suitable for dealing with special situations, such as occlusion, lighting changes, sudden movement and deformation. Some appearance-based tracking methods adopt either multimodal representation (Mahadevan et al., 2013) or nonlinear classifier (Chen et al., 2016) to separate the foreground from background and distinct co-occurring objects. However, the online model updating of these algorithms is limited to learning in time order, so it can't make the models fully differentiated and diversified.

The change of object appearance during tracking is one of the most difficult problems in visual tracking. It requires tracking algorithm to extract and learn new features quickly. The traditional tracking algorithm uses artificially designed low-level features (such as SIFT, HOG) to express targets. In recent years, deep neural networks (DNNs), especially convolutional neural networks (CNNs) (Lecun et al., 1998) has received extensive attention, CNNs had been widely used in online social networks
SUN et al. (2018), Internet (Sun et al., 2017) and other fields. For visual tracking, CNNs can extract high-level abstraction features that are more suitable for visual tracking tasks from digital images. Especially for color images, the convolution layer of CNN can directly extract features, which simplifies the process of feature extraction and data reconstruction compared with traditional features. Different from artificial extraction features, the features learned by CNNs from massive annotated visual data and a great deal of object classes include rich high-level semantic information and are strong at distinguishing objects of various categories. In addition, the neuron responses of CNN have strong selectiveness on object recognition, so those features are very robust to data corruption. Based on above advantages of CNNs, we decided apply CNNs to resolve the problems faced by tracking moving objects in large-scale video. The features of CNN at various levels have different characteristics which would fit the various kinds of tracking problem. Figure 1 is an example of a convolution neural network that can capture more abstract and high-level features.

**Figure 1.** Features of target localization based on CNN

In Figure 1 from left to right: input image, the truth target heat map, and the predicted heat maps using conv5-3 and conv4-3 layers of CNN respectively. The CNN can distinguish objects of different categories very clearly. But it can’t distinguish objects of the same category as shown in Figure 2.

**Figure 2.** Features of the same category based on CNN

Motivated by above fact, a novel CNN architecture is proposed in this paper. The new CNN offline learns the representation of targets from multiple annotated video sequences, and each video is regarded as a separate domain for obtaining generic image features. Then these features can be used for online tracking. The proposed online visual tracking algorithm consists of offline learning and online visual tracking, it can efficiently solve multi-modal object appearance problems and other special situations such as short-term occlusion and tracking failure. Therefore, more reliable models will be derived from online training with minimal training samples.

**Model**

In this section, the new CNN architecture and offline learning method for obtaining characteristic for visual tracking were described.

**The new CNN architecture**

The architecture of our proposed CNN is shown in Figure 3.

**Figure 3.** The architecture of our proposed CNN

There are shared layers (three convolutional layers (conv1-3) and two fully connected layers (fc4-5)) and $N$ branches of domain-specific layers ($fc6^{i}$-$fc6^{k}$) in the CNN. The size of the input image is equal to the size of the receiving space of a single unit (each domain) in the last layer. The output of the input image $x$ is the normalized vector $[\varphi(x), 1 - \varphi(x)]^T$. Its elements represent the points of the target and the background, respectively.

The early layers in our proposed CNN can keep more fine-grained information which can be useful for precise localization of moving objects. The visualized CNN characteristics on every convolutional layers of an image are shown in Figure 4.

**Figure 4.** Visualization of the CNN characteristics of an image
From Figure 4 we can see that the conv5-3 layer is less effective than the conv4-3 layer for precise localization.

The new CNN architecture can be managed by maintaining tree structure. In the tree structure $T = \{v, e\}$, the vertex $v$ corresponds to a CNN, the directed edge $(u, v)$ defines the relationship between CNN. The weight on the edge $(u, v)$ indicates the relationship between the two end points. It can be calculated as follows:

$$ s(u, v) = \frac{1}{|F_0|} \sum_{t \in F_0} \varphi_u(x_t^*) $$

(1)

where $F_0$ is a set of continuous frames for training CNN which is related to $v$. $x_t^*$ is the target state estimation at frame $t$. $\varphi_u(\cdot)$ is a positive predictive fraction of CNN relative to $u$.

The basic steps of the CNN model can be shown in Figure 5.

\[ Figure \ 5. \ \text{The \ basic \ steps \ of \ the \ proposed \ CNN \ model} \]

From the outputs of each convolutional layer, linear correlation can be obtained to infer the location of objects.

**Learning Algorithm**

The target of our learning algorithm is to train our proposed new CNN model to distinguish target and background accurately in large-scale video. The learning algorithm estimates the target state in the new frame by aggregating the scores of multiple CNN from the tree structure. By sampling the current target state, we will get the $N$ target candidates $x_1^*, x_2^* ...$ in the current frame $t$. The target score of sample $x_1^*$ is calculated from the weighted average of CNN.

$$ H(x_1^*) = \sum_{v \in TV} w_{v \rightarrow t} \varphi_v(x_1^*) $$

(2)

where $w_{v \rightarrow t}$ is the weight of the CNN corresponding to the target $v$ in the frame $t$, the best target $x_t^*$ is given by the candidate sample with the maximum value.

$$ x_t^* = \arg \max_{x_t^*} H(x_t^*) $$

(3)

The weight $w_{v \rightarrow t}$ is determined by the following two factors: the affinity of existing frameworks and the reliability of CNN. Affinity $a_{v \rightarrow t}$ represents how much of a CNN in $v$ affects the tracking results in frame $t$, it is determined by the maximum positive score of all the candidates.

$$ a_{v \rightarrow t} = \max_{x_t^*} \varphi_v(x_t^*) $$

(4)

Through this training process, object tracking information is modeled in the shared layers. A visualization example of convolutional layers is shown in Figure 6.

\[ Figure \ 6. \ \text{Visualization \ of \ convolutional \ layers} \]

As shown in Figure 6(d), characteristics on the fifth layer are effective in distinguishing the moving objects from background in video. So, the characteristic map selection method is very important. Our method is based on a target heat map regression model, which is trained by minimizing the square loss between the background heat map $B$ and the object heat map $O$.

$$ L_{set} = \|B - O\|^2 $$

(5)

The change of the loss function $\delta L_{set}$ can be computed as follows:
\[
\delta L_{sel} = \sum_i g_i \delta f_i + \frac{1}{2} \sum_i h_{ii}(\delta f_i)^2 + \frac{1}{2} \sum_{i\neq j} h_{ij} \delta f_i \delta f_j
\]  \hspace{1cm} (6)

where

\[
g_i = \frac{\partial L_{sel}}{\partial f_i}
\]  \hspace{1cm} (7)

\[
h_{ij} = \frac{\partial^2 L_{sel}}{\partial f_i \partial f_j}
\]  \hspace{1cm} (8)

So the significance of \(f_i\) can be defined as follows:

\[
s_i = -g_if_i + \frac{1}{2}h_{ii}f_i^2
\]  \hspace{1cm} (9)

**Target Localization**

After finishing characteristic map selection from the first frame, the SNet and the GNet were constructed by conv4-3 and conv5-3 characteristic maps, respectively, they are shown in Figure 7.

**Online Tracking**

The object to track is specified by target localization in current frame. In this paper tree structure management is used to store multiple target templates to ensure continuity and diversity of templates. The core idea of this tree structure is to establish a relationship between different templates to determine the dependence of different stages on different templates in the tracking process. The structure and working principles of the template tree in online tracking are shown in Figure 8.

After tracking the \(T\) frame continuous image \(F_x\), a new CNN is created at the node \(z\). Its parent CNN should make it most trustworthy, so the node \(p_z\) connecting the parent CNN can be given by the following formula:

\[
p_z = \arg \max_{v \in V_z} \min(s(v,z), \beta)
\]  \hspace{1cm} (13)

**Experiments**

In this paper we used Object Tracking Benchmark (OTB) and VOT2014 video databases to simulate the algorithm. Our algorithm is implemented in MATLAB and runs at a computer equipped with a 2.50 GHz Inteli5 CPU and NVIDIA GeForce GT555M GPU.

In experiment, we compared our algorithm with the eight classic trackers including MUSTer, CNN-SVM, MEEM, Struck, TGPR, DSST,
SCM and KCF. The precision and success plots of all compared algorithms are shown in Figure 9.

From Figure 9, we can see that our algorithm is much better than other trackers in both measures. Our algorithm can also effectively handle all types of abnormal situations. In particular, our algorithm can track moving objects in low-level characteristics of video. Four challenge attributes were chosen and the compared results were shown in Figure 10.

Figure 10. The results for four challenge attributes

Qualitative results of our algorithm compared with other classic trackers are shown in Figure 11.

Our algorithm was compared with other seven trackers including DSST, KCF, SAMF, MUSTer, DGT, MEEM, and PLT_14 on VOT2014 video database. The robustness and accuracy of all trackers were shown in Figure 12.

For some different challenge situations, the robustness and accuracy of all compared algorithms are shown in Figure 13.

From Figure 13, we can see that our algorithm is more stable than other trackers in different challenge situations.

Figure 11. Results of our algorithm on some challenging sequences

Figure 12. The robustness-accuracy plots of all trackers in VOT2014 dataset

Figure 13. The robustness-accuracy plots of all trackers for four challenge attributes
Conclusions
In this paper, an online visual object tracking method based on improved convolution neural network is proposed. Firstly, we used the improved convolutional networks to pre-train images and obtain characteristics for visual tracking. These characteristics can represent the semantic information of the moving object and distinguish the object from background in large-scale video. Through experiments, our algorithm is illustrated to greatly improve the tracking performance on challenging situations.

Acknowledgments
Authors gratefully thank the co-workers for concerning the research project and also acknowledge the support of Research Fund of Shaanxi Provincial Education Department (Grant No. 17JK0710), National Natural Science Foundation of China (Grant No.71501156) and Special Funds for Key Disciplines in Shaanxi Universities.

References